DOCUMENT RESUME

ED 419 836 TM 028 411

AUTHOR Gipps, Caroline; Clarke, Shirley; McCallum, Bet
TITLE The Role of Teachers in National Assessment in England.
PUB DATE 1998-04-00
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Elementary Secondary Education; Foreign Countries; *National Competency Tests; National Curriculum; Tables (Data); Teacher Attitudes; *Teacher Role; *Teachers; Test Use; Testing Problems
IDENTIFIERS *England

ABSTRACT

This paper explores the role of teachers in national assessment in England. The Education Reform Act of 1988 introduced a national curriculum for ages 5 to 16 together with a national assessment program for students at ages 7, 11, 14, and 16. The national assessment program is a crucial accompaniment to the national curriculum as the vehicle through which standards are to be raised. There are two main assessment methods, external tests or assessment tasks and teachers' own informal assessments of pupils' attainment called Teacher Assessment. For this, teachers make an assessment of each student's level of attainment on the scale of levels in relation to the attainment targets of the core subjects. Despite official support for the role of teachers in making assessment, there has been limited support for teachers to undertake this. In this context, teachers were surveyed about teacher assessment. In 1996, 288 teachers completed questionnaires, and 77 were interviewed. In 1997, 212 questionnaires were returned from teachers, and 216 from headteachers, and interviews were conducted with a teacher and the headteacher from 20 schools. Both studies made it clear that teachers think assessment is an essential process that has a direct impact on students' learning and their teaching. However, teachers still find the mechanisms, such as record keeping or standardization, time consuming. The studies also indicate that teachers use a variety of approaches in making their Teacher Assessment judgments, and that the assessment process would benefit from additional training for more standard approaches. (Contains 4 tables and 20 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 419 836

AERA Conference 1998

Symposium

Assessment Policy in the UK:

Issues for the Classroom and the Teacher

Convened by the British Educational Research Association

Assessment Policy Task Group

The Role of Teachers in National Assessment in England

Caroline Gipps, Shirley Clarke and Bet McCallum

Institute of Education, University of London

20 Bedford Way

London WC1H OAL, United Kingdom

Telephone: 0171 612 6315

email: c.gipps@ioe.ac.uk

TM028411

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Caroline Gipps

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

The Role of Teachers in National Assessment in England

Caroline Gipps, Shirley Clarke and Bet McCallum

Introduction

The United Kingdom is a small country, particularly in comparison with USA. It has four separate but linked areas: England, Scotland, Wales and Northern Ireland: a population of 58.4 million; and 4,455 secondary schools (age 11 - 16/15) and 23,408 primary schools (age 5 - 11). Compulsory schooling begins at age 5 and ends at age 16, although the majority of young people stay in education or training beyond this. Scotland and Northern Ireland have quite different education systems from that of England and Wales; England and Wales follow the same national curriculum and assessment structure, but in Wales the Welsh language is included as a first or second language. In this paper we shall concentrate on the developments and experience in England.

The National Curriculum

The Education Reform Act of 1988 introduced, for the first time in recent history, a national curriculum for age 5 - 16 together with a national assessment programme for pupils at ages 7, 11, 14 and 16.

The national curriculum was designed to ensure that all pupils of compulsory school age would follow the same course with English, mathematics and science forming the core, and history, geography, technology, a modern foreign language, art, music and physical education - the foundation subjects - forming an extended core. For each subject the curriculum is enshrined in law: statutory orders describe the matters, skills and processes to be taught as 'programmes of study' and the knowledge, skills and understanding as 'attainment targets' which pupils are expected to have reached at certain stages of schooling. The stages are

defined as Key Stage One (age 5 - 7), Two (age 7 - 11), Three (age 11 - 14) and Four (age 14 - 16).

The national assessment programme is a crucial accompaniment to the national curriculum for it is through the assessment programme that standards were to be raised; the first stage of the development of the national curriculum and assessment programme was the setting up of the Task Group on Assessment and Testing (TGAT). The report of this group (DES, 1988) put forward a blue-print for the structure of the curriculum to which all subjects had to adhere. Subjects are divided up into a number of components called attainment targets which are articulated at a series of progressive levels. The series of levels is designed to enable progression: most pupils of 7+ would be at level two in the system while most pupils of 11+ would be level four and so on. The attainment targets were described at each of the levels by a series of criteria or statements of attainment which formed the basic structure of a criterion-referenced assessment system.

There are two main assessment methods: external tests or assessment tasks; and teachers' own informal assessments of pupils' attainment called: Teacher Assessment (TA). For this, teachers make an assessment of each pupil's level of attainment on the scale of levels in relation to the attainment targets of the core subjects. Teachers may make these assessments in any way they wish, but observation, regular informal assessment and keeping examples of work, are all encouraged. The assessment tasks were originally designed as complex performance assessments with children engaged in active, curriculum-related tasks.

Because of the reliance on teacher assessment, the TGAT report suggested a complex process of group moderation through which teachers' assessments could be brought into line around a common standard. The combination of TA and test results has been a contentious area; the rule at first was that where an attainment target was assessed by both TA and test and the results differed, the test result was to be "preferred". Currently, the TA and test results are reported separately and they have equal weighting.

The first run of assessment for seven-year-olds in English, maths and science took place in 1991, with the second run in 1992. By the time 1993 came along the scope of the assessment programme and the time it took had been dramatically reduced, but many schools, as a result of a national boycott organized by the three largest teacher unions, either did not carry out the testing or did not report the results to their local education authorities (LEAs). In 1994 the boycott continued, but by 1995 all schools were carrying out national assessment at the end of Key Stage 1 and 2.

Boycott and Review

In 1992 and early 1993, there had been much debate amongst the teacher unions and professional associations as to whether teachers should boycott the testing. In March 1993, one teacher union declared a boycott on all national curriculum (NC) testing because of the perceived extra workload (approximately 116 hours) to carry out and mark the tests for 14 year olds. This was declared legal by the courts and other unions joined in. Although national curriculum assessment had largely settled down in infant schools many teachers of 7 year olds did not do the testing in 1993. As a result of the boycott the government set up a committee under the chairmanship of Sir Ron Dearing to review the entire national curriculum and assessment programme with the express aim that ways be found to simplify the testing programme.

In 1994 there was a second boycott by one teacher union only; but many primary school teachers either did not do the testing - a modified version for 7 year olds and the pilot tests for 11 year olds - or did the testing but did not report the results.

The major outcomes of the Dearing Review, were:

- a simplification of the curriculum

- the suspension of league tables of schools' performance at ages 7 and 14 (although the Government is committed to league tables of primary schools using 11 year olds results)
- the reporting of TA alongside test results and giving both equal status (rather than subsuming the teachers' assessments)
- a shift away from multiple statements of attainment to broad level descriptions.

Level Descriptions/Standards

In order to simplify the criterion referenced basis, almost one thousand statements of attainment have been reduced to 200 level descriptions. An example is given below:

> **Attainment Target 2: Number and Algebra**
>
> **Level 2**
>
> Pupils count sets of objects reliably, and use mental recall of addition and subtraction facts to 10. They have begun to understand the place value of each digit in a number and use this to order numbers up to 100. They choose the appropriate operation when solving addition and subtraction problems. They identify and use halves and quarters, such as half of a rectangle or a quarter of eight objects. They recognise sequences of numbers, including odd and even numbers.

These level descriptions are similar to the standards being used in parts of Australia and those being developed in the USA.

The move from statements of attainment to level descriptions has been made because of the overload provided by the huge number of statements of attainment within the core subjects of the national curriculum. However, one anxiety about the level descriptions is that they are too global to be used as assessment criteria, and that if teachers are to use them for

assessment purposes in anything more than a rough and intuitive way they may need to break them down; exemplars are also necessary in order to help classroom teachers make assessment against descriptions.

The Research

Despite official support for the role of teachers in making assessment within the national assessment programme there has been limited support for teachers to undertake this. Some training was given to Key Stage 1 teachers in the early stages of the programme but central funding for this has been withdrawn. Some 'non-statutory' assessment material has been provided to schools and this has proved popular. Exemplification materials, to support group and individual national judgements about levels of performance have also been produced (e.g. SCAA 1995).

Against this background in 1996 we undertook a research project for SCAA (The Schools' Curriculum and Assessment Authority), now known as QCA (the Qualifications and Curriculum Authority) to monitor the consistency of Teacher Assessment in England, across Key Stages 1, 2 and 3 and the extent of use of the centrally provided materials. That project was followed, in 1997, by another funded by the same authority to evaluate the 1997 national assessment at Key Stage 1, including both tests and Teacher Assessment.

For the consistency project (1996) data from a total of 288 questionnaires from Year 2 teachers (age 7), Year 6 teachers (age 11), Assessment Coordinators in primary schools and Heads of English, mathematics and science departments in secondary schools were analysed. Twenty four schools were visited as case studies and a total of 77 teachers were interviewed.

In the evaluation project (1997) a total of 212 questionnaires were returned from Year 2 teachers and 216 from headteachers; twenty schools were visited as case studies: the

headteacher and the Year 2 teacher were interviewed in each school and at least one test was observed.

In both projects we investigated how teachers make Teacher Assessment judgements and that data is brought together in this paper. Year 2 and Year 6 teachers were asked about their own practice. Heads of subject departments in secondary schools were asked about practice among subject specialist teachers in their departments.

There are three dimensions to making Teacher Assessment judgements in English schools:

a) **ongoing, day to day assessment**

b) **end of Key Stage (ages 7,11 and 14) TA judgements in terms of levels, to be reported for the core subjects**

c) **whole school or department standardisation meetings to ensure consistency of these TA level judgements**

Findings

We will present our data in relation to each of the three dimensions.

a) Ongoing, day-to-day assessment

In the 1996 consistency study teachers were asked to describe the elements of ongoing teacher assessment. A list of possible strategies was provided for teachers to tick as well as space to describe other strategies. Table 1 shows that primary teachers and English teachers in secondary schools have many aspects of their ongoing assessment practice in common. In general, it seems that mathematics and science departments in secondary schools adopt rather formal approaches to ongoing assessment (e.g. end of module test, regular classroom tests), whereas English departments and primary teachers tend to use more informal, formative methods (e.g. pupil self - assessment, regular notetaking, use of pupil portfolios).

Table 1

How teachers make ongoing day-to-day assessments

	PRIMARY		SECONDARY		
Elements of ongoing teacher assessment **N teachers**	**Year 2 teachers 60**	**Year 6 teachers 46**	**Head of English 34**	**Head of maths 31**	**Head of science 25**
Ongoing marking	56 - 93.3%	44 - 95.7%	33 - 97.1%	29 - 93.5%	21 - 84%
Regular informal assessments as part of the teaching plan	56 - 93.3%	42 - 91.3%	31 - 91.2%	26 - 83.9%	18 - 72%
Regular classroom tests	32 - 53.3%	37 - 80.4%	18 - 52.9%	26 - 83.9%	20 - 80%
Tracking significant achievement via a pupil portfolio	42 - 70%	30 - 65.2%	30 - 88.2%	15 - 48.4%	8 - 32%
Aspects of planning systems	44 - 73.3%	35 - 76.1%	19 - 55.9%	13 - 41.9%	10 - 40%
Involving pupils in self evaluation	29 - 48.3%	31 - 67.4%	29 - 85.3%	13 - 41.9%	13 - 52%
Regular collection of annotated samples of work	35 - 58.3%	23 - 50%	15 - 44.1%	10 - 32.3%	8 - 32%
Regular note - taking from structured or unstructured observations of practical and/or oral work	32 - 53.3%	28 - 60.9%	21 - 61.8%	5 - 16.1%	9 - 36%
Check lists based on level descriptions	30 - 50%	25 - 54.3%	14 - 41.2%	12 - 38.7%	7 - 28%
End of module tests with agreed criteria for the level to be awarded	15 - 25%	17 - 37%	9 - 26.5%	14 - 45.2%	20 - 80%

b) End of Key Stage TA level judgements

Evidence used to determine Teacher Assessment levels

The 1996 consistency study revealed that a variety of sources are used by teachers when deciding levels, as shown in Table 2. Most teachers, at this stage, said that the statutory test levels had no influence over Teacher Assessment levels (results are known before TA levels have to be completed).

Table 2

Evidence used to determine Teacher Assessment levels

	PRIMARY		SECONDARY		
Information used **N teachers**	**Year 2 teacher 60**	**Year 6 teacher 46**	**Head of English 34**	**Head of maths 31**	**Head of science 25**
General written work	58 - 96.7%	45 - 97.8%	33 - 97.1%	27 - 87.1%	22 - 88%
Set classroom tests or assessment activities	59 - 98.3%	40 - 87%	29 - 85.3%	30 - 96.8%	24 - 96%
Observations	59 - 98.3%	44 - 95.7%	29 - 85.3%	16 - 51.6%	20 - 80%
Dialogue with the pupil	54 -90%	40 - 87%	25 - 73.5%	17 - 54.8%	14 - 56%
The pupil portfolio	32 - 53.3%	26 - 56.5%	28 - 82.4%	13 - 41.9%	10 - 40%
Memory	40 - 66.7%	25 - 54.3%	12 - 35.3%	14 - 45.2%	10 - 40%
Homework	8 - 13.3%	10 - 21.7%	22 - 64.7%	21 - 67.7%	18 - 72%

Most teachers used general written work and regular classroom tests or assessment activities when deciding levels. Most teachers also used observations of pupils as a source of information, with the exception of heads of mathematics departments. Primary teachers and heads of English departments were more likely to consider dialogue with the pupil as a source of information, which may reflect the lack of opportunity for dialogue in mathematics and science departments. The pupil portfolio was used by around half of primary schools and particularly in English departments. Memory was more likely to be used by primary teachers, which may reflect the fact that primary teachers have the same class all year, so have a great deal of knowledge about their pupils. Understandably homework was used much more by secondary teachers as a source of information.

The 1997 KS1 evaluation study showed that the most common type of evidence used in Year 2 classes was teachers' records and children's work. Teachers' records of course, is a term likely to have many definitions, so we cannot say exactly what these records would consist of. Additional comments indicated that it is the more *recent* children's work which is used as evidence.

The process Attainment Targets (known as the AT1s) Speaking and Listening for English, problem solving for mathematics and science have been found in various SCAA evaluations to be the most difficult aspects of the curriculum both to teach and to assess. (Indeed, Section C of this paper describes how mathematics and science departments in secondary schools make AT1 the focus of all their standardisation meetings). The 1997 questionnaire asked separately about deciding TA levels for AT1; the data showed that, even as early as Year 2, teachers use specific assessment tasks in order to gather evidence for Attainment Target 1 for mathematics and science. For Speaking and Listening, however, the strategies are more widespread, including use of set tasks and discussion with colleagues, while memory is the most popular option.

Year 2 teachers were asked, in the questionnaire, whether they had used any of the SCAA test criteria to help them decide Teacher Assessment levels: 70.5% (148) said that they had. The task and test criteria (which are not the same as the level descriptions) are written in order to judge one piece of work, rather than for overall performance in an Attainment Target across a range of contexts. The use of task and test criteria in TA is a symptom of the lack of clarity in the level descriptions and of the 'best fit' approach. The writing task performance descriptions were most used (by 140 teachers), the main reason given that the criteria are much clearer than the Level Descriptions and because, as one teacher put it, *"They get into your consciousness"*.

The best fit approach

End of Key Stage TA level judgements are supposed to be based on the level descriptions from the 8 - level scale of the Attainment Targets for the various subjects. The statutory advice for determining a level is to apply a 'best fit' notion, which

> *"is based on knowledge of how the pupil performs across a range of contexts, takes into accounts strengths and weaknesses of the pupil's performance and is*

checked against adjacent level descriptions to ensure that the level awarded is the closest match to the child's performance in each attainment target"

QCA/DfEE 1998

The 1996 findings showed that most teachers did not think that the 'best fit' approach worked very well, because it was difficult to make decisions about pupils who appeared to fall between two levels and the notion of 'best fit' was too vague. However, having just been released from the previous system of counting the number of Statements of Attainment a pupil had attained in order to determine a level, teachers said that they found the approach much more manageable, so did not want it to be changed.

The 1997 evaluation found that, although Year 2 teachers still considered the approach manageable and did not want to return to the previous unmanageable system, another year's experience of making 'best fit' judgements had made them feel that it was not a good means of representing children's achievements. Questionnaire comments revealed that teachers felt that the approach was too open to different interpretations across schools.

It also emerged that teachers were dissatisfied with the 8 - level scale (graded level descriptions for each Attainment Target, intended to span the age ranges 6-14) in providing a continuum of performance. Although this is part of a complex picture, which includes the influence of league tables, it seems that secondary teachers tend not to believe the levels sent up to them by primary teachers, due to perceived generosity and lack of subject knowledge. Teachers also found it difficult to consider the levels without taking account of the age of the child, and the accompanying Programmes of Study for the age group. As one teacher put it:

" How can you relate a Level 3 five or six year old to a Level 3 15 year old? The disputes come from the structure itself: it means something different for different ages."

How teachers interpret 'best fit'

Bearing in mind that 1996 was the first year of determining TA levels in this way we attempted to find out exactly how teachers were interpreting and applying the concept of 'best fit'. A number of statements were given for teachers to tick if they agreed. As illustrated in Table 3 most teachers said that they made 'general best fit judgements'. Primary teachers and heads of English departments were more likely to use best fit judgements in relation to children's portfolios. (This links with the earlier findings about their ongoing assessment strategies). Approximately half of each group of teachers said that they identified key aspects of level descriptions (individuals must be able to do x, y and z in order to reach this level) in order to determine 'best fit', with the exception of heads of mathematics departments, where only 26% of teachers said that they did this.

Table 3

How teachers make 'best fit' judgements

	PRIMARY		SECONDARY		
N teachers	**Year 2 teachers 60**	**Year 6 teachers 46**	**Head of English 34**	**Head of mathematics 31**	**Head of science 25**
By making general 'best fit' judgements	43 - 71.7%	35 - 76.1%	18 - 52.9%	17 - 54.8%	18 - 72%
By using 'best fit' judgements in relation to children's portfolios	35 - 58.3%	22 - 47.8%	23 - 67.6%	10 - 32.3%%	5 - 20%
By splitting the level descriptions (eg. by creating separate statements and counting half or more as attaining a level)	12 - 20%	8 - 17.4%	4 - 11.8%	5 - 16.1%	3 - 12%
By identifying key aspects of level descriptions	31 - 51.7%	23 - 50%	14 - 41.2%	8 - 25.8%	13 - 52%

Interviewed teachers were also asked how they had used level descriptions to arrive at a level. Responses were very varied, but the overall picture was of secondary teachers averaging the

set of levels which pupils had by the end of the year and primary teachers using a general best fit judgement. We felt it would be interesting to pursue this further, to try to establish exactly how primary teachers define 'a general best fit judgement'. So more options were given to the Y2 teachers in 1997.

Table 4

How Y2 teachers interpret 'best fit' (1997) **N = 212**

'Best fit' interpreted as	**Yes**	**Case study sample N = 20**
The level description which overall describes the child's attainment better than the one above or below	71.7% (152)	45% (9)
Must achieve 75% or more of the statements in the level description	44.3% (94)	45% (9)
Must achieve important aspects of a level description	25.9% (55)	75% (15)
Intuition	17% (36)	35% (7)
Must achieve almost 100% or 100% of the statements in the level description	15.1% (32)	25% (5)
Must achieve 50% or more of the statements in the level description	1.9% (4)	0
Other	1.4% (3)	0

Table 4 shows that the most common interpretation of 'best fit' is to decide which level describes the attainment of the child more appropriately than adjacent levels. This statement was put in specifically at the request of SCAA officers and against our advice, since it does not tell us how the teacher makes the decision as to what is 'appropriate'. In order to decide that one level is more appropriate than another, some judgement has to be made, such as deciding key indicators or counting statements attained, or alternatively intuition.

Most teachers ticked more than one statement, indicating that a variety of strategies are used to make the generalised judgement, with the most common being 'counting 75% or more of the statements in a level achieved', followed by achieving important aspects of a level.

Interview responses from the Y2 case study teachers broadly supported the questionnaire responses: there are three strategies used (sometimes together) for deciding the level of attainment. Of the 20 teachers, 15 mentioned using key aspects of a level, 9 using 75% of the statements attained, 9 using an overall judgement checked against adjacent levels, seven using intuition; five said that they required 100% of the statements achieved (it seemed rather alarming that this group of teachers was applying the mastery principle). This variation across teachers is a concern, clearly, for consistency of judgements.

c) Whole school or department standardisation meetings

This topic was pursued through interviews in the 24 case study schools in 1996. Standardisation or consistency meetings (also known as group moderation or agreement trialling in the UK) are the main vehicle for enhancing consistency. However, the data revealed quite different approaches to these meetings between primary and secondary schools.

Secondary departments appeared to use standardisation meetings in order to check on marks awarded for school based tasks or tests, which were then used to determine the final level for a pupil. English departments tended to use marked samples of pupils' writing which were analysed against the level descriptions at the meeting, whereas mathematics and science departments set Attainment Target 1 tasks (investigations or problems) from three to six times a year then set up meetings specifically to check the grades awarded against levels. (It appeared that these were the only times when pupils encountered Attainment Target 1 work).

Primary schools used a range of pupils' work as the focus for standardisation meetings, analysing the work and deciding a school interpretation for the definition of a level. The work was then often put into a school portfolio, to be used for reference when deciding levels.

All teachers found the meetings, whether whole department or school, or of a small group, very useful and effective.

The 1997 evaluation indicated that primary schools were having fewer standardisation meetings of the kind described in the 1996 data, but were tending to focus more on standardisation meetings specifically to discuss borderline cases for the writing task in the statutory tests: at Key Stage 1 the teacher has to decide the test result, as these test papers are not externally marked. Thus it appears that the needs of the test have changed the focus of meetings, although Year 2 teachers' increasing familiarity with the level descriptions and use of 'best fit' has probably led to a perception that there is less need for standardisation meetings in which level interpretations are discussed.

Discussion

National Issues

A key element of national assessment policy is the involvement of teachers in assessing pupils' performance. This is important for two reasons: to give teachers a stake in the assessment process and to allow assessment of a broad range of skills and processes in order to maintain a broad curriculum. Teachers' comments about Teacher Assessment made it clear that they think it is an essential process which has a direct impact on pupils' learning and their teaching. However, teachers still find the mechanisms (e.g. record keeping, standardisation meetings) time consuming. Teachers indicated that, regardless of the workload problem, they wish to continue with Teacher Assessment in all its forms.

This research highlights the complexity of making what are essentially reporting judgements against broad descriptions of performance or performance standards.

The notion of 'best fit' is a consciously loose one. Because of this, teachers are taking a variety of approaches to making Teacher Assessment judgements. Some teachers will make quantitative judgements (to attain a level individuals must meet all the elements of a level description, 50%, or some other proportion); some will take a hurdle approach (individuals must be able to do x, y and z in order to reach Level 5); others will take an intuitive approach (this one feels like a good Level 4). Although not addressed in this study we know that some teachers will make ranking judgements (this individual is a clear Level 7, and this is a clear Level 6; less clear performances are then slotted in, in relation to these fixed points). Because of the lack of clarity of 'best fit', the differences in interpretation mean that, at times, there will have been a difference of one level awarded to pupils and this is not acceptable in a 'high stakes' programme.

Our findings point to the fact that agreement trialling, especially cross-school and cross-phase, is particularly important, as it is a crucial process to achieve consistency. Teachers clearly value Teacher Assessment and see its importance in maintaining a broad taught curriculum, they see standardisation meetings as valuable (despite the time issues); and primary teachers, especially, would value cross-school agreement trialling.

Other research has clearly shown that consistency in Teacher Assessment can best be achieved by use of exemplification materials **and** some form of group moderation (see Harlen, 1994). Ideally, there should be further 'Exemplification of Standards' publications, for all Key Stages, which provide more clarification and a greater range of examples of work, exemplifying: 'just reaches the level'; a 'safe level'; 'almost the next level'.

The differences in use of exemplification materials across primary and secondary are in large part due to the experience of the General Certificate of Secondary Education (GCSE) at secondary level which has involved secondary teachers in assessing and moderating pupils' work for many years. The examination boards which produce and mark these exams offer materials and procedures which are widely used. In our view, secondary mathematics and

science teachers need to be encouraged to use more informal, formative assessment methods, rather than relying on tests, so that more valuable feedback is provided to pupils. The most accessible of the strategies used by schools appears to be involving pupils in self assessment via target setting and the sharing of learning intentions. This resonates with good practice in learning.

The development of assessment skills among primary teachers shows how it is possible to give teachers a central role in any assessment programme. In 1990 much of their knowledge about assessment was rudimentary and their practice intuitive; tremendous development has taken place (Gipps et al, 1995; Brown et al 1997). But such development takes time, professional development and support material.

As a result of the reporting of TA in the high stakes context of performance tables (at age 11) there is some evidence (Woodhead, 1998) that the other functions of ongoing, day to day, assessment are being reduced: true formative assessment with feedback (Black and Wiliam, 1998) or what we call assessment **for** (as opposed to **of**) learning (Stobart and Gipps, 1997) seems to be less in evidence.

International issues

A number of countries have moved in recent years to implement national or state curriculum frameworks and assessment schemes which require teachers to report on student progress at designated times during primary and secondary schooling according to specified 'benchmarks' or 'standards' on various targeted learning outcomes. These standards are represented as developmental steps, stages or levels, in some cases based on national curriculum frameworks. In these schemes teachers play a key role in collecting evidence of student achievement and interpreting this evidence in terms of the specified performance standards. In some cases, teacher assessments may be supplemented by external tests. These schemes represent a substantial change from past educational practice, replacing the previous psychometric

paradigm of assessment, emphasising measurement, scaling and formal standardised tests, with the newer performance-standards paradigm, emphasising authentic and contextualised assessment and involving teacher judgement and interpretations of standards (Maxwell and Gipps, 1996). More is demanded of teachers in standards-based performance assessment. However, little is known about the actual processes of assessment. We hope that this English study goes some way to enhancing understanding.

In Australia (McGaw, 1996) there is a similar curriculum structure to the English one and an eight level progressive system of performance outcomes.

> "Curriculum development has long involved specification of scope and content and some form of declaration of the learning objectives held for students. What is new, in places like England and Australia at least, is an attempt to develop a more explicit standards perspective in curriculum by specifying student learning outcomes in developmental sequences.
>
> These sequences become specifications of standards when expectations of rates of student development are imposed on them as well. This can involve mapping of grade levels onto the sequences to indicate what some proportion of students is expected to have achieved by the end of each grade level or set of grade levels" (McGaw, 1996, p. 3).

As in the UK, consistency across teacher judgements is a major issue.

> "The outcome statements offer teachers a constant language for thinking about student learning and for discussing it with students and parents. It gives teachers the chance to use consistent criteria as a reference for student achievement.

> The question is can they use the criteria consistently with respect to other teachers or some group of experts, defined as 'experts' because they can make consistent, independent judgements of student performance against outcome scales?" (p. 13, ibid).

What McGaw's research has found is that teachers at different grade levels interpret the outcome statements in different ways: they make finer judgements in the range of outcomes in which their own students predominantly operate. In England, too, award of a level at different key stages is an issue with teachers of older ages generally not accepting that pupils from two or more years below can be operating at the level appropriate for the older key stages. We found that the majority of teachers were in favour of the eight level scale, but felt that the levels were too broad. (The eight levels cover nine years of school in England and Wales; ten years of school in Australia). Many of our teachers said that the eight level scale was **not** being used as a continuum because of the difficulties of comparing pupils with the same level across different key stages and because of the hiatus which occurs when pupils transfer from one school to another: pupils may be 'stuck' for a number of years because the levels in the previous school are not seen as realistic by the receiving school.

In New Zealand, the concept of 'sufficiency' has been researched (Keown 1996) in relation to assessment of pupils against unit standards. The standards have a pass/fail structure and the research was concerned to prevent secondary teachers from doing too much formal assessment and to encourage them to build on 'naturally occurring' evidence instead.

> "Quality sufficiency decision making can be defined as a process of collecting the quantity and quality of evidence required to convince an assessor that a candidate is or is not competent in relation to the function defined by the element without over assessing or under assessing". (Keown, 1996, p. 3).

In the Keown study teachers' main concerns were: the quality of response required to award credit; quantity of evidence required to award credit; the problem of time for assessment and reassessment; how to collect naturally occurring evidence fairly; authenticity and how to cope with group work and homework; how to track record and feed back evidence and performance quickly enough for students to be able to benefit and present better at a re-sit; and how to get consistency between teachers and between schools. The author concluded that there is a need to institute a programme of training to assist teachers to broaden their repertoire of assessment strategies so that they can gather valid naturally occurring evidence to supplement their formal assessment activity evidence, thus reducing the amount of formal assessment required.

However, the use of naturally occurring evidence is not so simple, as the findings from our study indicate. If the use of assessment results is high stakes then reliability and consistency issues come in to full play. *Consistency of standards* relates to ensuring that different teachers interpret the assessment criteria in the same way, whether using naturally occurring evidence or setting tests. However, where tests are used it may be necessary to ensure *consistency of approach*: the assessment task or activity which is used and the way in which such tasks are presented to the pupil, or indeed contextualised, can affect performance quite markedly. To ensure consistency of approach, therefore, we need to ensure that teachers understand fully the constructs which they are assessing (and therefore what sort of tasks to set); how to get at the pupil's knowledge and understanding (and therefore what sort of questions to ask); and how to elicit the pupil's best performance (the physical, social and intellectual context in which the assessment takes place). This, of course, is a tall order.

Group moderation is a key element of teacher assessment, not only in terms of improving inter-marker reliability, but to support the *process* of assessment too. If we wish to be able to 'warrant assessment-based conclusions' without resorting to highly standardised procedures with all that this implies for poor validity, then we must ensure that teachers have

common understandings of the criterion performance and the circumstances and contexts which elicit best performance: this can be developed through group moderation.

The disadvantage of group moderation is that it is time consuming and costly and this may then be seen to add to any unmanageability in an assessment programme. Its great advantage, on the other hand, lies in its effect on teachers' practice (Linn, 1993; Radnor and Shaw, 1994). It has been found that where teachers come together to discuss performance standards, or criteria, the moderation process becomes a process of teacher development with wash-back on teaching. It seems that coming together to discuss performance or scoring is less personally and professionally threatening than discussing, for example, pedagogy. But discussion of assessment does not end there: issues of production of work follow on and this broadens the scope of discussion and impacts on teaching (Gipps, 1994, p. 80).

It is possible, and, we would argue, desirable to give teachers a role in an assessment programme but the underlying requirements are complex. What this paper explores are the issues around evidence, judgement and consistency in such teacher assessment practice.

Bibliography

Black, P. and Wiliam, D. (1998 in press) Assessment and Classroom Learning, *Assessment in Education*, Vol. 5, no. 1, pp. 7-73.

Brown, M., McCallum, B., Taggart, B. and Gipps, C. (1997) The Validity of National Testing at age 11: the Teacher's View, *Assessment in Education*, Vol. 4, no. 2, pp. 271-293 .

Clarke, S. and Gipps, C. (1997) *Evaluation of Key Stage 1 Statutory Assessment (England) 1997*, SCAA, London.

DES (1988) *National Curriculum: Task Group on Assessment and Testing: A Report*, DES/Welsh Office.

Gipps, C. (1994) Quality in Teacher Assessment, in Harlen, W. (ed.), *Enhancing Quality in Assessment*, Paul Chapman, London. pp. 71-86.

Gipps, C. and Clarke, S. (1996) *Monitoring Consistency in Teacher Assessment and the Impact of SCAA's Guidance Materials at Key Stages 1, 2 and 3*, SCAA, London.

Gipps, C., Brown, M., McCallum, B. and McAlister, S. (1995) *Intuition or Evidence? Teachers and National Assessment of Seven Year Olds*. Open University Press, Milton Keynes.

Harlen, W. (Ed) (1994) *Enhancing Quality in Assessment*. BERA Policy Task Group on Assessment, Paul Chapman Publishers.

Keown, P. (1996) *Walking the Sufficiency Ridge*. Final Report of the Sufficiency Issue Research Project, University of Waikato.

Linn, R. L. (1993) Linking Results of Distinct Assessments. *Applied Measurement in Education*, Vol. 6, no. 1, pp. 83-102.

Maxwell, G. and Gipps, C. (1996) Teacher Assessments of Performance Standards: A Cross-National Study of Teacher Judgements of Student Achievement in the Context of National Assessment Schemes. *Application for funding to the ARC: Interdisciplinary and International Research.*

McGaw, B. (1996) *Technical Issues in Assessments*. paper to the American Educational Research Association Annual Meeting, New York, April 1996.

QCA/DfEE (1995) *Assessment and Reporting Arrangements 1998* QCA Publications, London.

Radnor, H. and Shaw, K. (1994) Developing a collaborative approach to moderation: The moderation and assessment project-southwest, in Torrance, H. (ed.), *Evaluating Authentic Assessment*, Open University Press, Buckingham.

SCAA (1995) *Consistency in Teacher Assessment, Key Stages 1 to 3*, SCAA, London.

SCAA (1995) *Exemplifcation of Standards. English Key Stages 1 and 2. Levels 1 to 5.* SCAA. London.

SCAA (1995) *Exemplification of Standards. Mathematics. Key Stages 1 and 2. Levels 1 to 5.* SCAA. London.

SCAA (1995) *Exemplification of Standards. Science. Key Stages 1 and 2. Levels 1 to 5.* SCAA. London.

Stobart. G. and Gipps. C. (1997) *Assessment: A Teachers' Guide to the Issues.* Hodder and Stoughton. London. (3rd edition).

Woodhead. C. (1998) *Blood on the Tracks: Lessons from the History of Education Reform.* HMCI Annual Lecture. Ofsted. London.